NYT ETL Pipeline

This project implements an ETL pipeline to fetch and analyze articles from the New York Times API.

Project Structure

```
NYT/
├── .env
├── .git/
├── .gitignore
├── README.md
├── requirements.txt
├── run.py
├── app/
│   ├── __pycache__/
│   ├── templates/
│   │   ├── base.html
│   │   ├── data.html
│   │   ├── etl.html
│   │   └── index.html
│   ├── __init__.py
│   └── routes.py
├── data/
│   ├── articles.db
│   ├── common_words_YYYY-MM.csv
│   └── headline_analysis_YYYY-MM.csv
├── logs/
│   └── data_fetch.log
├── notebooks/
│   └── view_all_articles.ipynb
├── scripts/
│   ├── __pycache__/
│   ├── analytics.py
│   ├── drop_all_tables.py
│   └── word_count_headline_analysis/
│       ├── analyze_headlines.py
│       ├── articles.py
│       ├── bar_chart.py
│       ├── bar_chart_html.py
│       ├── generate_wordcloud.py
│       ├── generate_wordcloud_html.py
│       ├── run_etl.py
│       └── transform_headlines.py
├── sql/
│   ├── articles_by_news_desk.sql
│   ├── articles_by_source.sql
│   ├── articles_by_type.sql
│   ├── articles_by_word_count_range.sql
│   ├── articles_per_month.sql
```

```
│   ├── avg_word_count_by_source.sql
│   ├── common_words_plot_last_month.sql
│   └── top_headlines.sql
└── static/
    ├── css/
    │   └── styles.css
    └── plots/
        ├── common_words_cloud_YYYY-MM.html
        ├── common_words_cloud_YYYY-MM.png
        ├── common_words_plot_YYYY-MM.html
        └── common_words_plot_YYYY-MM.png
```

Getting Started

Prerequisites

- Python 3.9 or higher
- `pip` package installer

Installation

1. Clone the repository:

```
git clone https://github.com/mkonefal2/NYT.git
cd NYT
```

2. Create and activate a virtual environment:

```
python -m venv venv
source venv/bin/activate  # On Windows use `venv\Scripts\activate`
```

3. Install the dependencies:

```
pip install -r requirements.txt
```

4. Create a `.env` file in the root directory and add your API key:

```
NYT_API_KEY=your_actual_api_key_here
```

Usage

To run the ETL pipeline for last month, execute the following script:

```
python  ./scripts/word_count_headline_analysis/run_etl.py
```

Output will be visible in :

```
./static/plots/
```

You can also run web app

```
python -m streamlit run .\scripts\analytics.py
```